Supplemental Material for "Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data"

Jean Fan[1,*], Hae-Ock Lee[2,*], Soohyun Lee[1], Da-eun Ryu[2], Semin Lee[1], Catherine Xue[1], Seok Jin Kim[5], Kihyun Kim[5], Nikolaos Barkas[1], Peter J. Park[1], Woong-Yang Park[2,§], Peter V. Kharchenko[1,3,4,§]

A. Supplemental Methods

B. Supplemental Figures

C. Supplemental References

A. Supplemental Methods

Generating allele count matrices

Putative heterozygous SNP sets from ExAC hg19 are available in the HoneyBADGER package and functions for generating allele count matrices are provided for both cases where cells are provided as individual bams as well as when cells are merged into the same bam file but differentiated by cell barcodes, as is common for newer droplet-based sequencing protocols. For each putative heterozygous SNP position, the read counts for the reference and alternate allele as well as the total read depth were summed across all cells belonging to a single patient. Any SNPs for which there was no coverage in any cells for a patient, or for which the sum of reference and alternate allele reads did not equal the total read depth were removed prior to downstream analyses.

Single cell differential expression analysis

MM34 and MM34A cells were clustered by hierarchical clustering using the ward.D method on Euclidean distance for posterior probabilities of CNVs. MM34 BM-like and extramedullary-like subclones were identified by cutting the dendrogram in two.

Differential expression analysis on the two identified subclones was performed using SCDE (v1.99.1) (Kharchenko et al. 2014; Fan et al. 2016) with default parameters following recommended protocols (http://hms-dbmi.github.io/scde/diffexp.html).

Significantly differentially expressed genes were identified using an absolute non-corrected Z-score cut-off of 1.96, corresponding to p-value < 0.05, for heatmap visualization, and 1.28, corresponding to p-value < 0.2, for gene set enrichment analysis.

Gene set enrichment analysis was performed using the LIGER (https://github.com/JEFworks/liger) package with input values as sorted MLE estimates of fold-change limited to significantly differentially expressed genes. In total, 10593 curated (C2), GO (C5), oncogenic (C6), and immune (C7) gene sets from MSigDB (Liberzon et al. 2015) were tested. Gene sets with less than 5 genes or more than 500 genes were omitted.

Estimation of mono-allelic detection probability

Mono-allelic detection rate was estimated by analyzing the heterozygous SNPs as identified by the bulk WES analysis of the MM34 sample. SNPs were stratified into 10 coverage quantiles and the average fraction of heterozygous SNPs for which only one allele was detected was calculated for each bin.

Estimation of effective sequencing error

To estimate the rate of effective sequencing error in SNP detection resulting from PCR, RT, RNA-editing, or other artifacts, we took heterozygous SNPs from known copy number neutral regions identified by bulk WES, and assessed the rate of detection for neither annotated allele. To accommodate for the under-estimation due to misinterpretation of sequencing errors as true heterozygous variants, the resulting rate estimate was doubled. The effective sequencing error rate is used in both the emission model for the allele-based HMM and the allele-based Bayesian hierarchical model. We estimated this effective sequencing error for MM16 and MM34, for which we had previous WES data, to be approximately 0.01. Due to lack of WES for additional datasets analyzed, we continued to use this effective sequencing error rate of 0.01. However, we acknowledge that with different sequencing platforms or library construction techniques, this effective sequencing error may differ and may need to be re-estimated for optimal performance.

Filtering out putative RNA-editing

Comparing pooled frequency of alleles across single cells to what's observed in WES, if the observed deviation is too improbable based on a binomial distribution

$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} < 0.05$$

where p is the alt-allele fraction observed in WES
n is the pooled coverage across all single cells
and k is the pooled alt-allele count across all single cells

bias may have been introduced by RNA-editing and thus sites are removed from downstream analysis. Applied to MM16, this filtering approach removed 1138 out of 8280 (14%) of SNPs. Where bulk WES was not available and potential RNA-editing is not removed, estimation of effective sequencing errors may be increased to accommodate.

Hidden Markov model – extended

HoneyBADGER contains implementations of an expression-based HMM as well as an allele-based HMM to identify regions potentially affected by CNVs. HMM were implemented using the HiddenMarkov package in R.

For the expression-based HMM, a transition matrix is defined on 3 hidden states representing deletion, neutral, and amplification $\begin{bmatrix} 1-2t & t & t \\ t & 1-2t & t \\ t & t & 1-2t \end{bmatrix}$ where t=1e-5 by default. Emission probabilities are defined by a normal distribution with means and variance estimated from the normalized expression data (see *Expression-based approach*).

For the allele-based HMM, a transition matrix is defined on 2 hidden states representing deletion or LOH, and neutral $\begin{bmatrix} 1-t & t \\ t & 1-t \end{bmatrix}$ where t=1e-5 by default. Recall that we define the lesser allele as the allele that is less frequently observed across our population of cells. In the presence of a subclonal CNV, we would expect to see persistent depletion of this lesser allele across our population of cells, either due to expression being only from the non-deleted allele for a deletion. Intuitively, the probability of observing the lesser allele in the neutral diploid case should be approximately p=0.5, as there is equal likelihood of detecting either allele. However, by the definition of the lesser allele, it must be observed with a probability less than 0.5 on a population scale. Based on testing known neutral diploid regions of multiple datasets, we find the rate of observation of the lesser allele in the neutral diploid case to occur at p=0.45. In the event of a deletion or LOH, emission from the lesser allele could only occur due to sequencing error, PCR errors, unfiltered RNA-editing, other artefacts, or contamination by cells not harboring the deletion or LOH being tested. We have tested multiple datasets and have determined this effective error rate to be approximately 0.1. Therefore, the emission probabilities are defined by a binomial distribution with the size parameter given by the pooled coverage at the SNP position and an expected p=0.1 for the lesser allele in the case of deletion or LOH and p=0.45 for neutral.

Default transition probabilities transition have been set based on the size of the regions expected to be able to detect. Benchmarking the impact of transition probabilities using MGH31, which harbors a known deletion on Chr 10, and MM135, which harbors a known deletion of Chr 13. Indeed, as t increases, we observe a higher false positive rate as expected. Yet in retesting each candidate CNV, we find that the false positive CNVs exhibit very low posterior probability in all cells, resulting in the final correct identification. A substantial decrease in the provided transition probabilities will eventually lead to more false negatives, as expected. Using t=1e-13, our HMM identifies no CNVs. Generally, we find that both the expression-based HMM and allele-based HMM is robust to choices of t from 0.001 to 1e-8 (Supplemental Fig. 9) in the limited number of cancer samples, genomic regions, and single-cell protocols we have tested. However, for genomic regions and protocols with high rates of erroneous SNP detection or high normalized expression variance due to technical noise, we anticipate that these transition probabilities may need to be tuned.

Hierarchical Bayesian model – extended

HoneyBADGER contains implementations of an expression-based approach, an allele-based approach, ad an integrative approach for assessing the posterior probability of CNVs in given regions. All Bayesian hierarchical models were written in BUGs for Gibbs sampling. Simulation from the models using MCMC was accomplished through rJAGS. Four chains were initialized specifying starting values for S^k and dd^k as 0 or 1 in all possible permutations where appropriate. The MCMC chains were allowed to run for 1000 iterations, with an adaptation of 100 and a burn-in of 100. Trace plots were used to ensure appropriate mixing on the hyper parameters and Gelman plots were used to diagnose convergence of chains (Supplemental Fig. 10A, B).

Expression-based approach:

For a particular region of interest, our goal is to make inference on the copy number status of a cell for that region given its observed gene expression in the region relative to a putative diploid expression reference of comparable cell type. For a candidate region, we assume that the observed average normalized gene expression is reflective of the underlying, hidden copy number status. If there is copy number loss, we expect there to be an expression magnitude change by $-\theta_0$ on average, even though due to complex regulatory dynamics, individual genes within a deletion may still exhibit increases in expression relative to a normal reference. Likewise, for a copy number gain, we expect there to be an expression magnitude change by $+\theta_0$ on average relative to a normal reference. For copy number neutral regions, we expect no shift in the expression magnitude on average compared to a normal reference. Genes used for modeling should be restricted to only highly expressed genes in both the test and normal expression references to better ensure that these deviations in expression are reflective on average of underlying copy number changes rather than due to differences in drop-out rates, which are substantially higher for single cell data (Kharchenko et al. 2014; Islam et al. 2014; Patel et al. 2014). The expression magnitude for each individual gene can be variable, as captured by ϵ, but the general trend should be consistent with the underlying copy number variation status. For a candidate region, let $S^k = 1$ if cell k has a copy number variation and $S^k = 0$ if cell k is copy number neutral. Let $dd^k = 1$ for a copy number gain and $dd^k = 0$ for a copy number loss, such that S^k and dd^k together capture the copy number status for cell k. Let $\overline{gexp^k}$ be the observed average normalized gene expression for genes within the tested region of interest in cell k. Thus, we seek to estimate the posterior distribution of $\left[\, S^k, dd^k \mid \overline{gexp^k} \,\right]$. We can accomplish this through a hierarchical Bayesian framework, modeling the observed gene expression as a function of the variables of interest: $\left[\, S^k, dd^k \mid \overline{gexp^k} \,\right] \propto \left[\, \overline{gexp^k} \mid S^k, dd^k \,\right]\left[\, S^k, dd^k \,\right] = \left[\, \overline{gexp^k} \mid S^k, dd^k \,\right]\left[\, S^k \right]\left[\, dd^k \,\right]$

The components of the Bayesian hierarchical model are as follows:

Data Model: $\overline{gexp^k} = \theta_k + \epsilon_k \,, \epsilon_k \sim Normal\left(0, \sigma_k^2\right)$

Process Model: $\theta_k = \begin{cases} \theta_0 & if\ S^k = 1, dd^k = 1 \\ 0 & \quad if\ S^k = 0 \\ -\theta_0 & if\ S^k = 1, dd^k = 0 \end{cases}$

Parameter Model: $S^k \sim Bernoulli\ (\alpha_k)\ , dd^k \sim Bernoulli\ (\beta)$

Where σ_k^2 is a cell-specific variance to account for expected noise around $\overline{gexp^k}$ as an estimate of θ_k. β is a hyper-parameter shared across cells. Thus, we assume if there is a copy number variation within a region, the direction of the variation will be the same for all cells and therefore a particular region cannot simultaneously harbor both deletions and amplifications in the same population of cells. α^k is cell-specific. θ_0, σ_k^2 are estimated from each dataset. Given a region of known copy number alteration, we estimated θ_0, σ^2 by randomly sampling gene sets of variable size and computing their mean. Both the mean value and variance about the mean were dependent on the size of the gene set (Supp. Fig. 10C, D). Therefore, we fit a curve for the expected variance about that mean.

The BUGs model representation is as follows:

```
model {
  ## Single cell
  for ( k in 1:K ) {
    for ( j in 1:JJ ) {
      ## Expression level
      gexp[j, k] ~ dnorm (mu[k], sigma0[k])
    }
    # 0 if neutral
    # + mag if amplification
    # - mag if deletion
    mu[k] <- 0 * ( 1 - S[k] ) +
         -mag0 * ( S[k] * (1 - dd)) +
         mag0 * ( S[k] * dd)

    ## Cell level
    S[k] ~ dbern (alpha[k]) # cnv or not
    alpha[k] ~ dunif (0,1) # cell specific hyper-parameter
  }
  dd ~ dbern (beta) # direction of cnv
  beta ~ dunif (0,1) # cell specific hyper-parameter
}
```

Allele-based model:
For a particular region of interest, our goal is to make inference on the copy number variation status of a cell for that region given its observed allelic imbalance for germline heterozygous SNPs within the region. For a candidate region, we assume that the allelic ratios are reflective of the underlying, hidden copy number status. If there is copy number loss, we expect to only be able to see expression from the non-deleted allele.

For heterozygous SNP i in gene j of cell k, let r_{ij}^k be the observed alternative allele read count and n_{ij}^k be the observed total read depth at that SNP site. Here, for a candidate region, let $S^k = 1$ if cell k has a deletion (or LOH) and $S^k = 0$ if cell k is copy number neutral. Our goal is to make inference on S^k. If bulk tumor whole exome-seq is available, for heterozygous SNP i in gene j, let l_{ij} be the observed alternative allele read count and m_{ij} be the observed total read depth at that SNP site. Alternatively, single cells may be pooled and putative heterozygous SNP sites be inferred from SNP databases such as ExAC. For heterozygous SNP i in gene j, let $a_{ij} = 1$ if the alternative allele is the deleted allele and $a_{ij} = 0$ if the reference allele is the deleted allele. We make inference on a given our observations, l_{ij} and m_{ij}. Intuitive, if we see depletion of the alternative allele as evidenced by $\frac{l_{ij}}{m_{ij}} < 0.5$, then we would have reason to believe $a_{ij} = 1$. Let b_j^k be a Bernoulli random variable indicating whether gene j in cell k is has mono-allelic detection. b_j^k is a function of the expected mono-allelic detection rate derived from the data. Let $d_j^k = 0$ if mono-allelic detection is for the reference allele and $d_j^k = 1$ if mono-allelic detection is for the alternative allele. Let e be a constant error rate estimated from the data to take into consideration sequencing error and RNA-editing. A *pseudo* parameter is used to help with mixing.

It should be noted that inference of a_{ij} establishes phasing. For example, if the first SNP in gene j $a_{1j} = 1$ and the second snp in gene j $a_{2j} = 1$, it must be that both reference alleles are on the same allele. We do not anticipate that explicit phasing of SNPs from scRNA-seq data alone will result in substantially improved performance, as such explicit phasing would be limited to nearby expressed heterozygous SNPs, impacting less than 20% of all heterozygous SNPs, with only 5% of local phasing impacting more than 2 SNPs based on estimates from our SmartSeq2 data. However, where long-range phasing information is available through integration with WGS or additional datasets, a_{ij} may be fixed a priori rather than inferred from observed data.

The BUGs model representation is as follows:

```
model {
  ## Single cell
  for ( k in 1:K ) {
    ## Gene level
    for ( j in 1:J ) {
      ## SNP level
      for ( i in 1:I.j[j] ) {
        r[j,i,k] ~ dbin (p[j,i,k], n.sc[j,i,k])
        ## if deletion, bias has no effect since only one allele to express
        ## the one allele expressed should be consistent with bulk
        ## if no deletion, there could be mono-allelic expression
        ## which affects deviation away from expected 0.5 probability
        ## b is imputed from our bias model
        p[j,i,k] <- ( h[j,i,k]* (1-b[j,k]) +
                       (pseudo*d[j,k] +
                       (1-pseudo)* (1-d[j,k]))*b[j,k] )* (1-S[k]) + fma[j,i]*S[k]
        h[j,i,k] ~ dnorm (0.5,0.1)T (pseudo, 1-pseudo) ## heterozygous snp prob
      }
      ## Need to fully define in order to pull out from model even if not used
      for ( i in I.j[j]+1:max (I.j) ) {
        p[j,i,k] <- 0
        h[j,i,k] <- 0
      }
      d[j,k] ~ dbern (0.5) ## random direction of bias
      b[j,k] ~ dbern (mono) ## probability of mono-allelic expression
    }
    S[k] ~ dbern (alpha[k])
    alpha[k] ~ dunif (0,1) # cell specific hyper-parameter prior to allow for better mixing
  }
  ## Bulk
  for ( j in 1:J ) {
    for ( i in 1:I.j[j] ) {
      l[j,i] ~ dbin (fma[j,i], n.bulk[j,i]) # lesser allele count
      # prob of observing lesser allele
      fma[j,i] <- pseudo* (ma[j,i]) + (1-pseudo)* (1-ma[j,i])
      ma[j,i] ~ dbern (0.5) # whether lesser allele is affected, each is independent
    }
    ## Need to fully define in order to pull out from model even if not used
    for ( i in I.j[j]+1:max (I.j) ) {
      fma[j,i] <- 0
      ma[j,i] <- 0
    }
  }
}
```

Combined model:

The BUGs model representation is as follows:

```
model {
  ## Single cell
  for ( k in 1:K ) {
    ## Gene level
    for ( j in 1:J ) {
      ## SNP level
      for ( i in 1:I.j[j] ) {
        r[j,i,k] ~ dbin (p[j,i,k], n.sc[j,i,k])
        p[j,i,k] <- (S[k] * (1-dd)) * fma[j,i] +
         ( 1 - (S[k] * (1-dd)) ) * (
         ( 1 - b[j,k] ) * h[j,i,k] +
               b[j,k] * ( pseudo * d[j,k] + ( 1 - pseudo ) * ( 1 - d[j,k] ) )
        )
        h[j,i,k] ~ dnorm (0.5,0.1)T (pseudo, 1-pseudo)
      }
      for ( i in I.j[j]+1:max (I.j) ) {
        p[j,i,k] <- 0
        h[j,i,k] <- 0
      }
      d[j,k] ~ dbern (0.5) ## random direction of bias
      delta[j,k] ~ dunif (0,1) ## random degree of bias
      b[j,k] ~ dbern (mono) ## probability of mono-allelic expression
    }
    ## Gene level
    for ( j in 1:JJ ) {
      ## Expression level
      gexp[j, k] ~ dnorm (mu[j,k], sigma0[k])
      # 0 if neutral
      # + mag if amplification
      # - mag if deletion
      mu[j,k] <- 0 * ( 1 - S[k] ) +
        -mag0 * ( S[k] * (1 - dd) ) +
        mag0 * ( S[k] * dd )
    }
    S[k] ~ dbern (alpha[k]) # cnv or not
    alpha[k] ~ dunif (0,1) # cell specific hyper-parameter
  }

  ## Bulk
  for ( j in 1:J ) {
    for ( i in 1:I.j[j] ) {
      l[j,i] ~ dbin (fma[j,i], n.bulk[j,i])
      fma[j,i] <- pseudo* (ma[j,i]) + (1-pseudo)* (1-ma[j,i])
      ma[j,i] ~ dbern (0.5)
    }
    for ( i in I.j[j]+1:max (I.j) ) {
      fma[j,i] <- 0
      ma[j,i] <- 0
    }
  }

  dd ~ dbern (beta) # direction of cnv
  beta ~ dunif (0,1) # cell specific hyper-parameter
}
```

Scalability of hierarchical Bayesian model:

Runtime scales linearly with respect to the number of SNPs or genes with a fixed number of cells, and likewise linearly with respect to the number of cells with a fixed number of SNPs or genes. Based on simulated benchmarks by sampling variable numbers of SNPs, genes, and cells, we approximate that for a single tested CNV using a linear fit:

runtime in seconds = 0.08795 * ncells

runtime in seconds for 75 cells = 1.48617 + 0.08328*(ngenes or nsnps)

on a single core. Additional complexities in subclonal structure resulting more candidate CNVs being tested will also add to the runtime. Computation can be distributed across multiple cores, thus improving scalability when high number of cores are available.

MM135 10X Genomics Sample collection and analysis

Bone marrow aspirates were obtained for clinical testing and an aliquot was cryopreserved in freezing medium (90% FBS and 10% dimethylsulphoxide; Sigma Life Science, St. Louis, MO, USA) following red blood cell lysis. After rapid thawing and washing in complete medium (Iscove's modified Dulbeco's media containing 10% FBS), cells were resuspended in 1X PBS as 1*10^6 cells/ml. Cellular suspension was loaded on a GemCode Single-Cell Instrument (10x Genomics, Pleasanton, CA, USA) targeting 5000 cells and single-cell RNA-Seq libraries were prepared using GemCode Single-cell 3' Gel Bead and Library kit following the manufacturer's instructions. Sequencing was performed on an Illumina HiSeq2500 with 2x100bp paired-end mode targeting 20G output per sample.

Alignment, gene expression quantification, and cell filtering was performed using the CellRanger v1.3 pipeline from 10X Genomics, resulting in 1340 cells. For expression-based clustering, gene expression counts were first normalized to counts per million (CPMs). We derived 100 principal components (PCs) with the largest eigenvalues based on the 1000 most highly variable genes after variance normalization. Infomap graph-based clustering with k=30 on the 100 PCs was used to identify transcriptional subpopulations. Visualization was achieved using tSNE with perplexity=30.

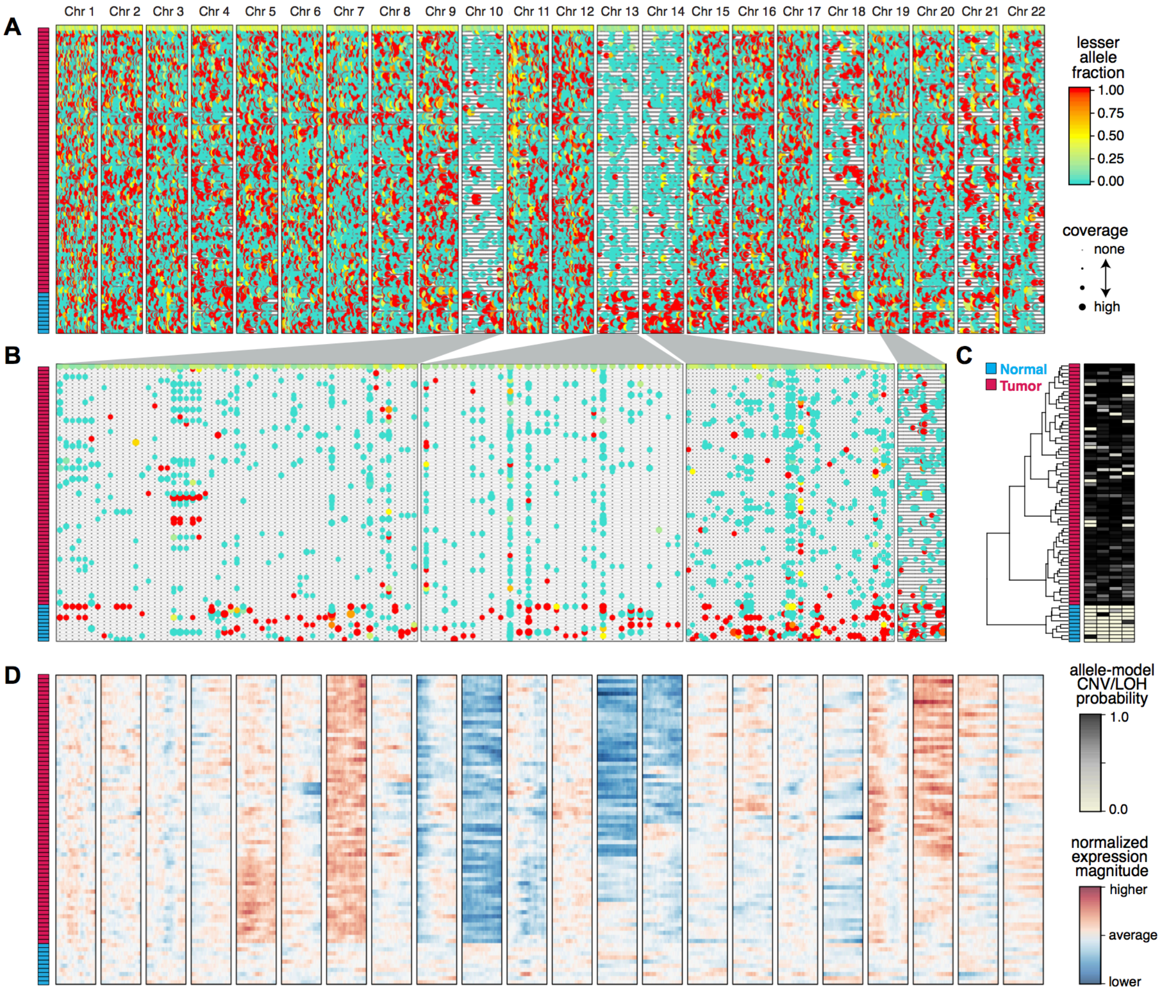
To identify putative heterozygous SNPs, common SNPs were identified from ExAC and assessed for coverage across all cells as described previously. The allele-based approach was first applied to identify a set of 135 putative normal cells based on a high posterior probability of lack of alterations. Expression profiles from these putative normal cells were averaged to establish a normal gene-expression reference. Both the allele and expression-based HMMs were applied to identify all potential CNVs. The combined Bayesian hierarchical model was applied to derive posterior probabilities for each CNV in each cell. Resulting CNVs with greater than 0.55 posterior probability in more than 100 cells were used for hierarchical clustering with the ward.D method.

Cytogenetic analysis was performed using standard G-banding after short-term culture of the bone marrow aspirates. Chromosomal abnormalities were described according to the Inernational System for Human Cytogenetic Nomenclature 2009. Fluorescence in situ hybridization was performed on $CD138^{+}$ sorted cells, using specific probes for IGH/CCND1 t(11;14), IGH/FGFR3 t(4;14), IGH/MAF t(14;16), CSK1B(1q21), IGH (14q32), D13S319(13q34), P53 (17p13.1), and CEP17 (D17Z1) (Abbott Molecular probes, Des Plains, IL, USA) in the Department of Laboratory Medicine and Genetics, Samsung Medical Center.

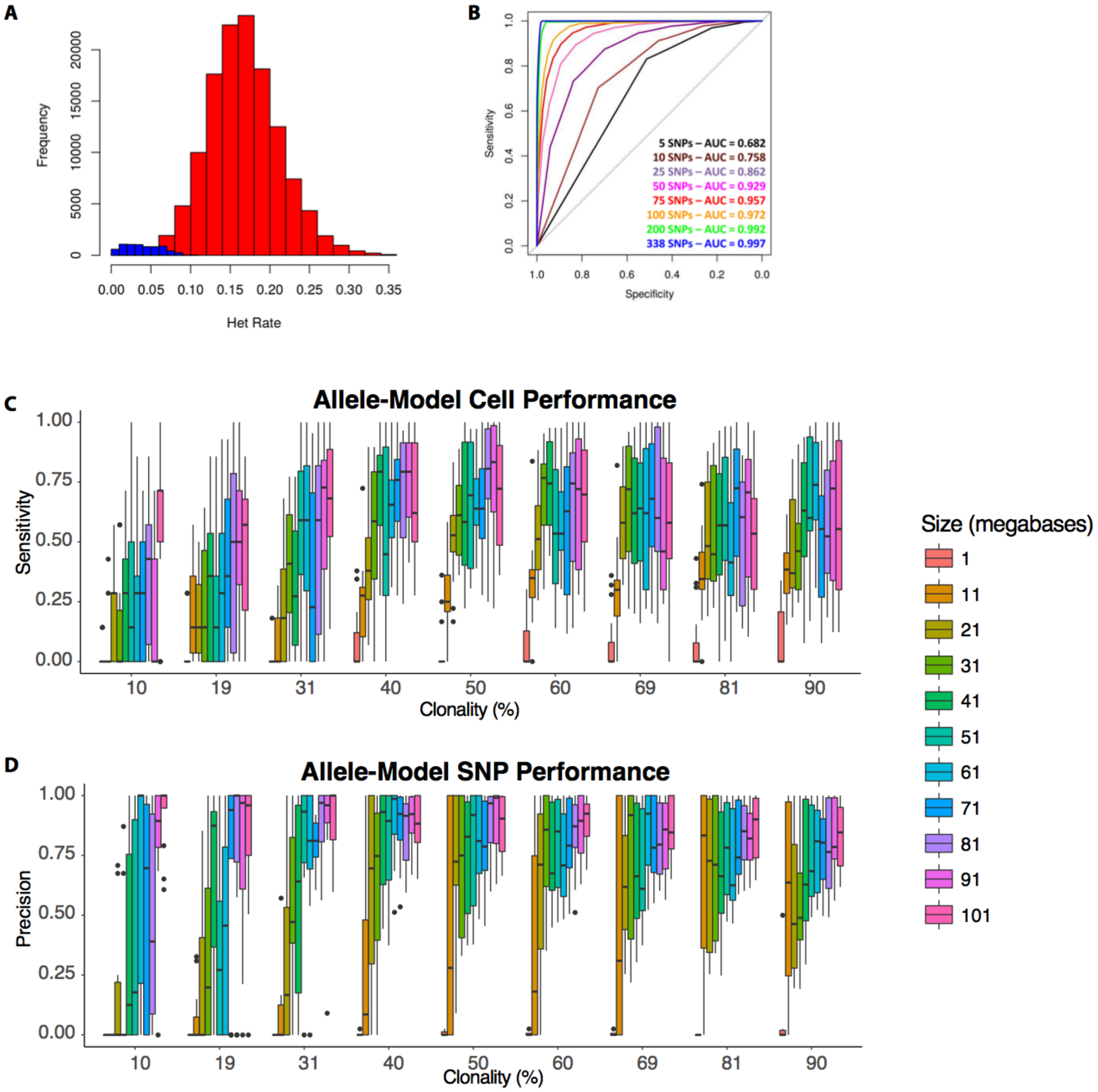
Simulating smaller CNVs

Focal deletions of varying size were simulated by matching lesser allele fractions or gene expression from known neutral regions to known deletion regions in tumor cells. To simulate across various regions, we randomly select start sites from 0 to 1e8 by 1e7 bases. We vary the simulated deletion region from 1Mb to 101Mbs by 10 Mbs. We then match the number of SNPs or number of genes within the region to SNPs and gene expression from known deletion regions. Posterior probabilities for deletion status from HoneyBADGER are binarized into discrete deletion calls at various thresholds to compute the accuracy and precision of deletion calls.

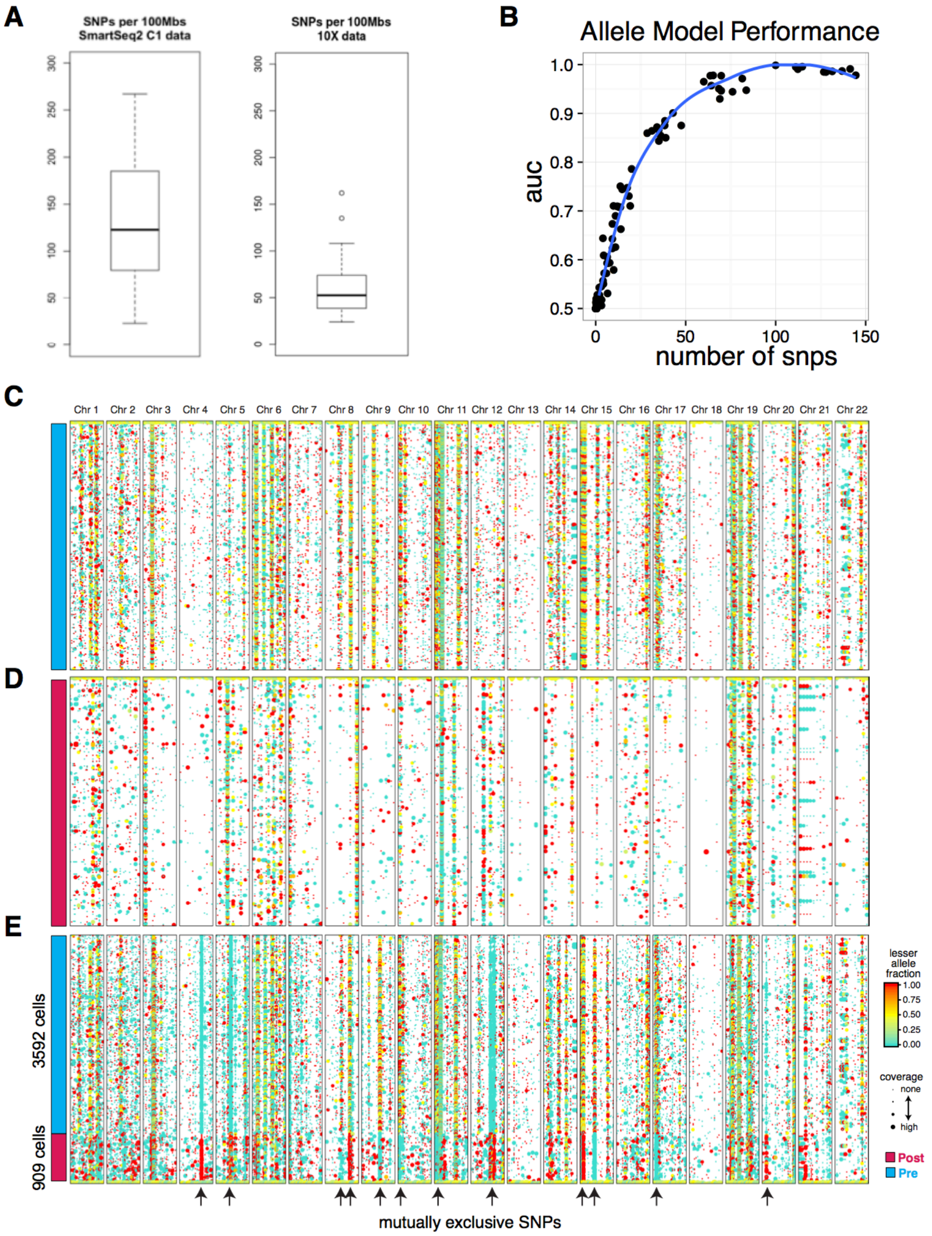
B. Supplemental Figures



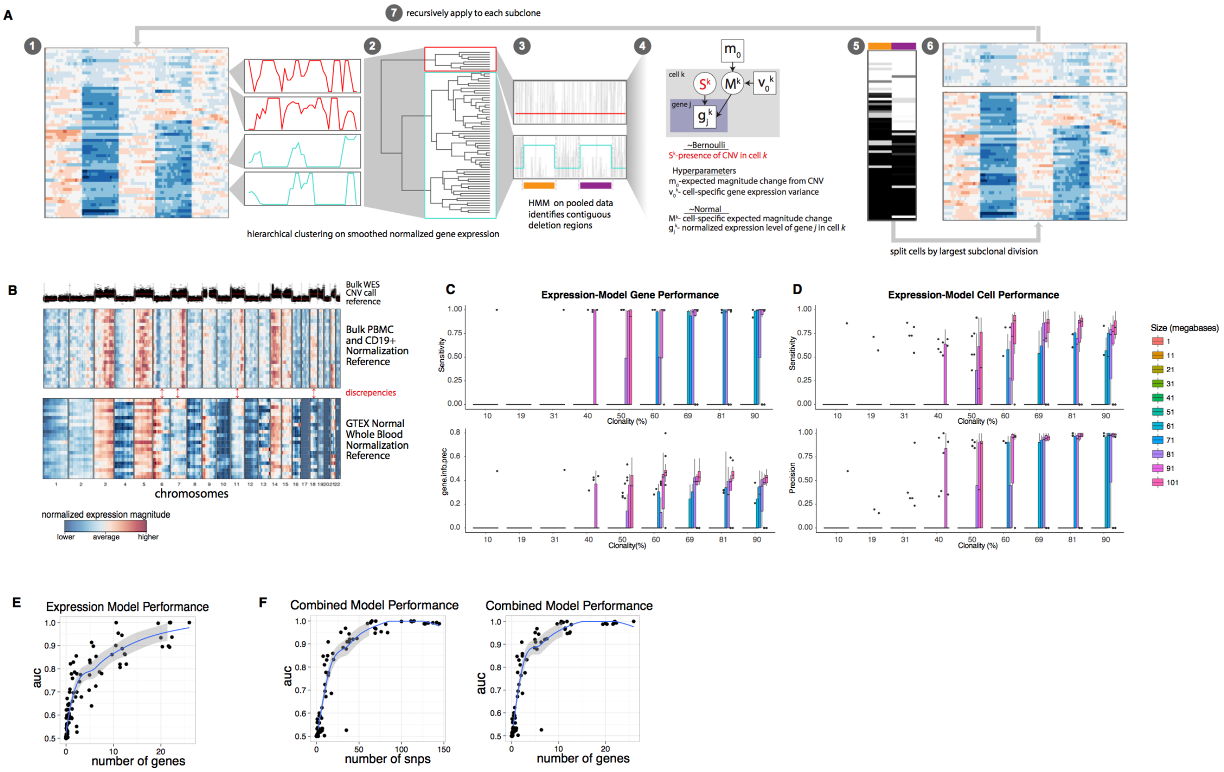
Supplemental Figure S1. Application of HoneyBADGER to published GBM MGH31 dataset from Patel et al. A) Lesser allele profiles where each column is a heterozygous SNP and each row is a single cell. Points are colored by the lesser allele fraction with yellow suggesting equal detection of both alleles and red and blue indicated mono-allelic detection in either direction. Points are sized by coverage at the SNP site in the given cell. Cells are ordered based on row dendrogram in C. **B)** Allele profiles for regions identified by HoneyBADGER as potential CNV or LOH regions. Width corresponds to size of region. Cells are ordered based on row dendrogram in C. **C)** Heatmap of posterior probability of CNVs or LOHs in each identified region where each column is a region and each row is a cell. Row side colors annotate cells classified as Normal or Tumor. **D)** Normalized gene expression profiles for 75 cells from patient MGH31 recapitulates the original Figureure 1B from Patel *et al*. Normalization reference of bulk normal brain samples from GTEx used as in the original publication.



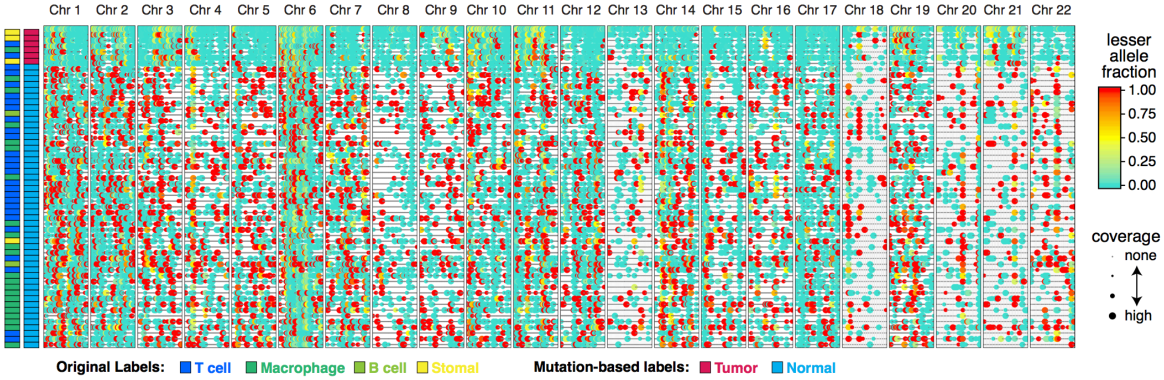
Supplemental Figure S2. Additional HoneyBADGER allele-model performance benchmarks. A) Rate of detecting common heterozygous SNPs from ExAC for a known clonal chromosomal deletion (blue) and neutral regions (red) based on scRNA-seq data from patient MGH31. Substantially fewer common heterozygous SNPs can be detected in clonal chromosomal deletion regions as expected due to the presence of only one allele. Detection of heterozygous SNPs in clonal chromosomal deletion regions may be interpreted as background from sequencing errors. **B)** Simulations demonstrate the ability to detect smaller clonal deletions based on depleted rate of common heterozygous SNPs. For windows of 5, 10, 25, 50, 75, 100, 200, or 338 consecutive common heterozygous SNPs, we assess the ability to distinguish between known deletion and neutral regions based on the rate of detected heterozygous SNPs in MGH31. **C)** Allele-model precision for identifying SNPs affected by deletion. **D)** Allele-model sensitivity for distinguishing tumor - cells with deletion - from normal - cells without. **E)** Allele-model as a function of the number of SNPs in a candidate CNV region.



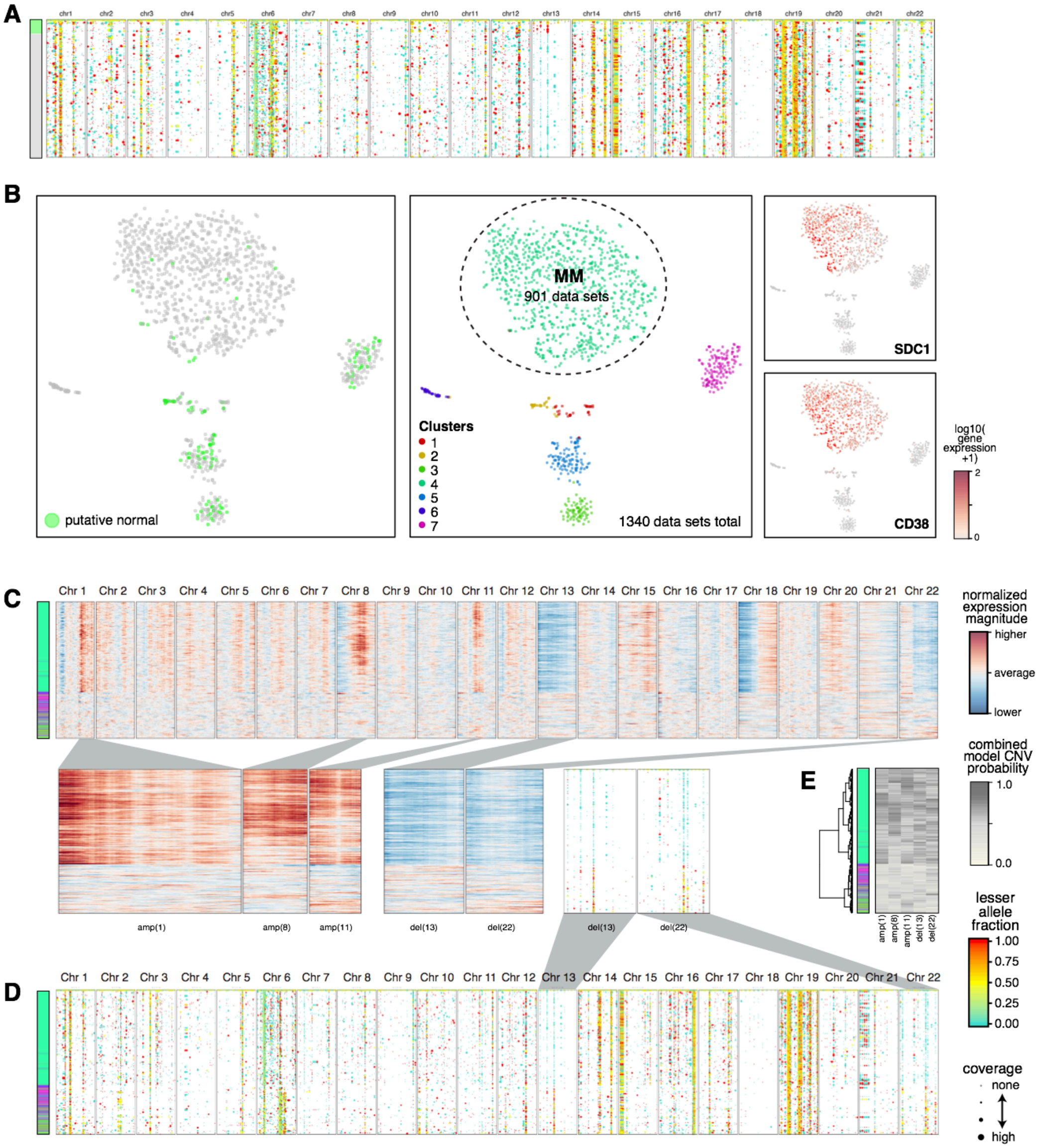
Supplemental Figure S3. Application of HoneyBADGER to published 10X genomics AML035 dataset from Zheng et al. A) Heterozygous SNP density comparison between full-transcript SmartSeq2 C1 data from Patel et al. and 10X Genomics 3' sequencing data from Zheng et al. Heterozygous SNPs were identified based on detected common heterozygous variants from ExAC in pooled single cells (Methods) where 75 cells were pooled from MGH31 from Patel et al. and 3592 cells were pooled from AML035-pre from Zheng et al. Although transcript coverage is substantially reduced with 3' sequencing, our ability to identify heterozygous SNPs within these regions is substantially improved due to the high number of cells**. B)** Alelle-model as a function of the number of SNPs in a candidate CNV region. **C)** Lesser allele profile for AML035-pre. **D)** Lesser allele profile for AML035-post. **E)** Joint analysis of AML035-pre and post samples identifies numerous mutually exclusive SNPs with those at high coverage across multiple cells highlighted by arrows.



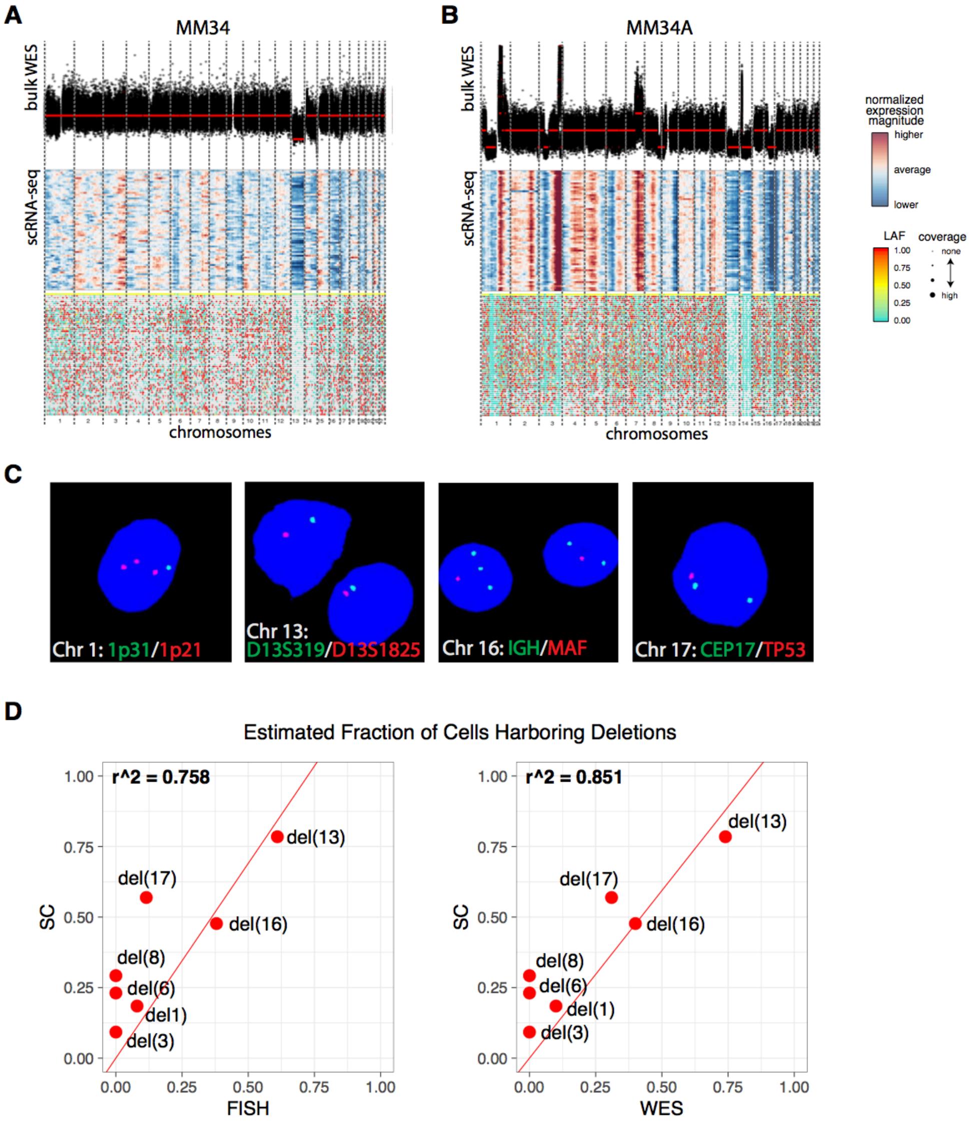
Supplemental Figure S4. Overview of expression-based model in HoneyBADGER and performance. A) CNVs are identified from scRNA-seq data in the following 7 steps. (1) Cells are first clustered on smoothed normalized gene expression profiles. (2) Cells are split into 2 main groups and pooled (3) A hidden markov model on the pooled normalized gene expression profiles identified with potential CNVs (4) A Bayesian hierarchical model assessed the posterior probability of a deletion or amplification for each region in each cell (5) Cells are clustered by their posterior probabilities of deletion or amplification for each region (6) Cells are split into putative subclones (7) Approach is recursively applied to each subclone until no new subclones can be detected. **B)** Variability in normalized expression profiles depending on normalization reference. Normalization reference of bulk normal PBMCs and CD19^{+} cells (top) and the normal blood samples from GTEx (bottom). Discrepancies are highlighted with red arrows. Notably, a deletion appears present on chromosome 6 with the bulk PBMC and CD19^{+} normalization reference, but not with the GTEx Normal Whole Blood normalization reference. Similarly, a deletion incorrectly appears present on chromosome 19 with the GTEx Normal Whole Blood normalization reference but not with the bulk PBMC and CD19^{+} normalization reference, thus highlighting the importance of finding an appropriate normalization reference for such an expression-based approach**. C)** Expression-model sensitivity and precision for identifying genes affected by deletion. **D)** Expression-model sensitivity and precision for distinguishing tumor - cells with deletion - from normal - cells without. **E)** Expression-model as a function of the number of genes in a candidate CNV region. **F)** Combined model as a function of the number of SNPs and genes in a candidate CNV region.



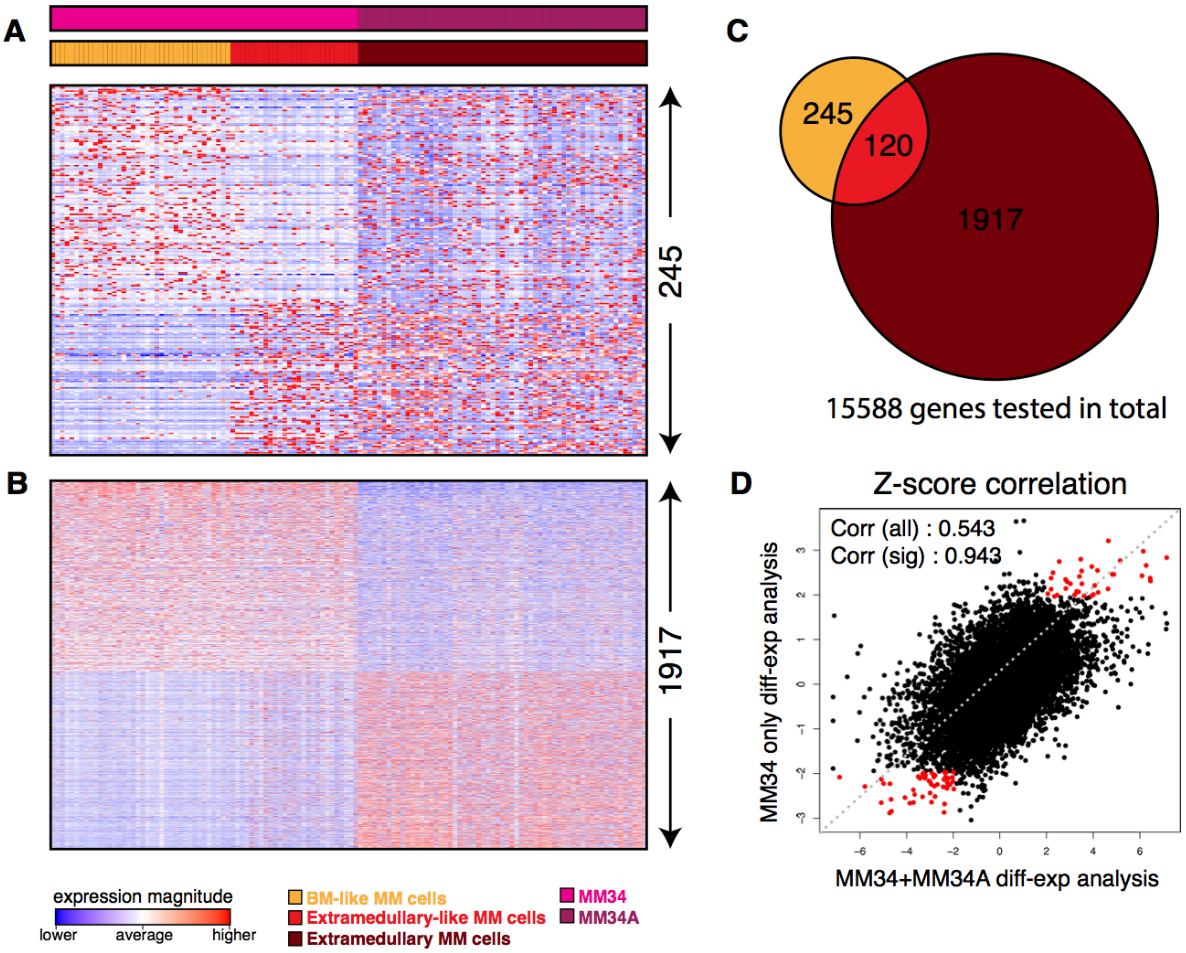
Supplemental Figure S5. Application of HoneyBADGER to published breast cancer BC09 dataset from Chung et al. Lesser allele profiles where each column is a heterozygous SNP and each row is a single cell. Points are colored by the lesser allele fraction with yellow suggesting equal detection of both alleles and red and blue indicated mono-allelic detection in either direction. Points are sized by coverage at the SNP site in the given cell. Column side colors annotate cells by their inferred labels from expression-based karyotyping from the original publication, as well as based on presence of breast cancer-related point mutations, which is consistent with annotations from HoneyBADGER.



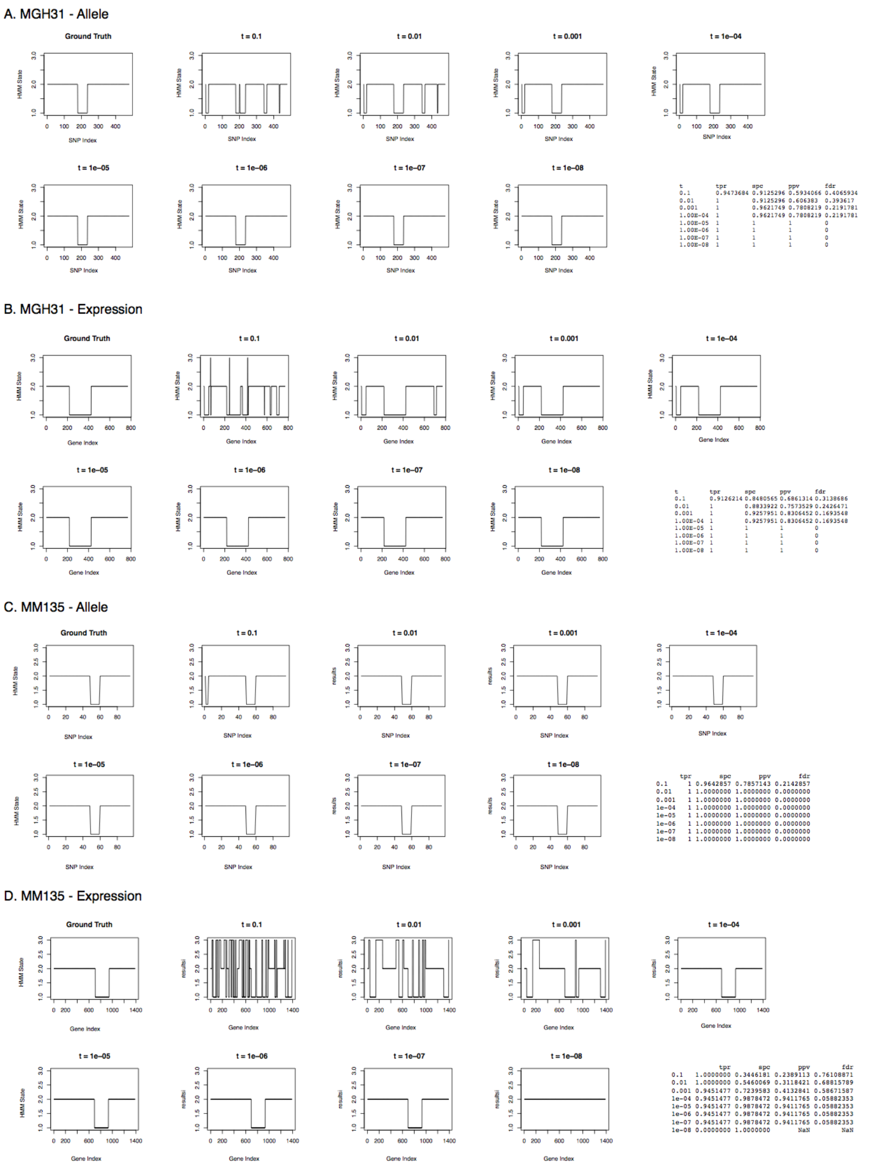
Supplemental Figure S6. Application of HoneyBADGER to 10X genomics MM135 dataset. A) Allele-based approach identified a subset of putative normal cells. **B)** tSNE embedding of 1340 unsorted bone marrow cells visualized transcriptional subpopulations and cell types. Putative normal cells identified from A are visualized in green. Clusters are identified using Infomap graph-based clustering and colored accordingly in the embedding. Expression of MM marker genes SDC1 (CD138) and CD38 are shown on the right, with red being high expression and grey being no detected expression. Both MM markers localized to one cluster, which has been circled and labeled as MM cells. **C)** Normalized gene expression profile for all cells using the average of non-MM cells as the expression reference. Cells are ordered based on the clustering shown in E. Column side colors label cells according to their cluster annotation from B. Normalized gene expression for all identified CNV regions and lesser allele profiles for all identified deletions below. **D)** Lesser allele profiles for all cells. Again, cells are ordered based on the clustering shown in E. Column side colors label cells according to their cluster annotation from B. **E)** Heatmap of posterior probabilities of identified CNVs with row dendrogram from hierarchical clustering.



Supplemental Figure S7. Genetic characterization of MM cells from the MM patient 34 collected at two distinct time points. A) Analysis for the initial bone marrow sample MM34. Bulk WES CNV calls identify deletion on chromosome 13 (top). Normalized gene expression profiles for the individual cells (rows) against a normalization reference of bulk normal PBMCs and CD19$^+$ cells (middle). Lesser allele frequency (LAF) profiles for single cells (bottom). **B)** Analysis of a sequential ascites dissemination sample (MM34A). **C)** Interphase FISH cytogenetics of MM34. Representative fields of view are shown. Of the 200 interphase cells from MM34 analyzed, 8.0% had a single CDKN2C(1p32) and three CKS1B (1q21) signals, 61.0% had a single D13S319(13q14.3), D13S1825 (13q34) signal, 9.5%-12.5% had three IGH signals, 38.0% had a single MAF signal, and 11.5% had a single p53(17p13.1) signal. A separate analysis estimates the sample to have 83-96% tumor purity by CD138+. **D)** HoneyBADGER predictions are highly consistent with FISH and bulk WES estimates. Fraction of single cells harboring deletions based on HoneyBADGER predictions (posterior probability > 80%) exhibits statistically significant correlation with the estimates from FISH (p-value = 0.0158) and bulk WES (p-value = 0.00216).



Supplemental Figure S8. Transcriptional characterization of genetic subclones. A) MM34-only transcriptomic analysis. Differential expression analysis of the MM34A-like subclone versus other MM34 cells identifies 245 significant differentially expressed genes (p-value < 0.05). **B)** MM34 and MM34A joint transcriptomic analysis. Joint comparisons of the MM34A-like subclone with MM34A cells versus other MM34 cells identifies 1917 significantly differentially expressed genes (p-value < 0.05). However, sample or batch specific effects may be observed. **C)** Overlap of significantly differentially expressed genes. Out of 15588 genes tested, 120 are consistent between analysis from A and B. **D)** Correlation of significance Z-scores between differentially expressed genes between MM34-only and joint analysis. In general, we see a strong correlation, 0.543 for all genes and 0.943 for the 120 jointly significant genes, suggesting consistency of findings irrespective of the significance threshold used.



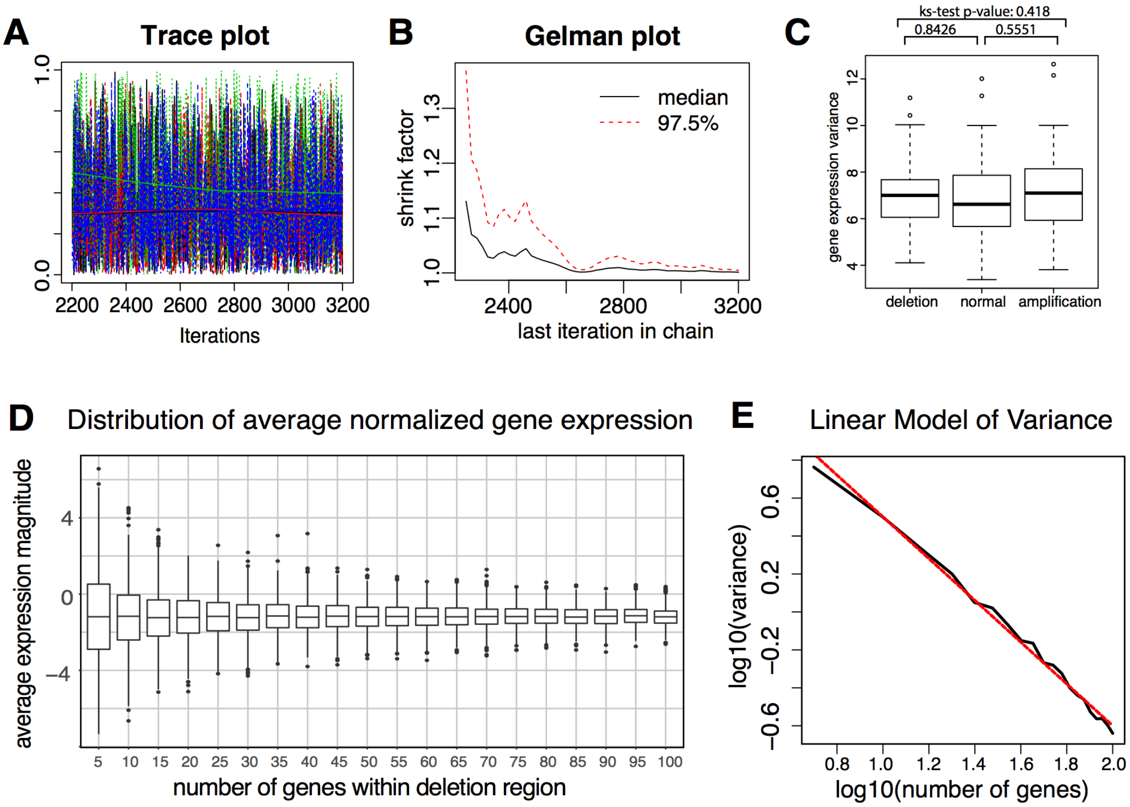
t	tpr	spc	ppv	fdr
0.1	0.9473684	0.9125296	0.5934066	0.4065934
0.01	1	0.9125296	0.606383	0.393617
0.001	1	0.9621749	0.7808219	0.2191781
1.00E-04	1	0.9621749	0.7808219	0.2191781
1.00E-05	1	1	1	0
1.00E-06	1	1	1	0
1.00E-07	1	1	1	0
1.00E-08	1	1	1	0

t	tpr	spc	ppv	fdr
0.1	0.9126214	0.8480565	0.6861314	0.3138686
0.01	1	0.8833922	0.7573529	0.2426471
0.001	1	0.9257951	0.8306452	0.1693548
1.00E-04	1	0.9257951	0.8306452	0.1693548
1.00E-05	1	1	1	0
1.00E-06	1	1	1	0
1.00E-07	1	1	1	0
1.00E-08	1	1	1	0

	tpr	spc	ppv	fdr
0.1	1	0.9642857	0.7857143	0.2142857
0.01	1	1.0000000	1.0000000	0.0000000
0.001	1	1.0000000	1.0000000	0.0000000
1e-04	1	1.0000000	1.0000000	0.0000000
1e-05	1	1.0000000	1.0000000	0.0000000
1e-06	1	1.0000000	1.0000000	0.0000000
1e-07	1	1.0000000	1.0000000	0.0000000
1e-08	1	1.0000000	1.0000000	0.0000000

	tpr	spc	ppv	fdr
0.1	1.0000000	0.3446181	0.2389113	0.76108871
0.01	1.0000000	0.5460069	0.3118421	0.68815789
0.001	0.9451477	0.7239583	0.4132841	0.58671587
1e-04	0.9451477	0.9878472	0.9411765	0.05882353
1e-05	0.9451477	0.9878472	0.9411765	0.05882353
1e-06	0.9451477	0.9878472	0.9411765	0.05882353
1e-07	0.9451477	0.9878472	0.9411765	0.05882353
1e-08	0.0000000	1.0000000	NaN	NaN

Supplementary Figureure S9. Robustness of HMM transition probabilities. A) For our allele-based HMM, we test the ability to identify a known clonal Chr 10 deletion (ground truth) using allele data from Chr 9, 10, and 11 from MGH31, and plot the HMM inferred states for various parameter values for t used in the transition matrix: $\begin{bmatrix} 1-\mathrm{t} & \mathrm{t} \\ \mathrm{t} & 1-\mathrm{t} \end{bmatrix}$. A table (bottom right) is also provided to summarize the true positive rate (tpr), specificity (spc), positive predictive value (ppr) and false detection rate (fdr). **B)** For our expression-based HMM, we test the ability to identify a known clonal Chr 10 deletion (ground truth) using normalized gene expression data from Chr 9, 10, and 11 from MGH31, and plot the HMM inferred states for various parameter values for t used in the transition matrix: $\begin{bmatrix} 1-2\mathrm{t} & \mathrm{t} & \mathrm{t} \\ \mathrm{t} & 1-2\mathrm{t} & \mathrm{t} \\ \mathrm{t} & \mathrm{t} & 1-2\mathrm{t} \end{bmatrix}$. Likewise, a summary table is provided (bottom right). C and D) Allele-based HMM and expression-based HMM respectively to identify a known clonal Chr13 deletion (ground truth) using data from Chr 12, 13, and 14 from MGH31.



Supplemental Figure S10. MCMC diagnostics and model parameter estimation. A) Trace plots show appropriate exploration of parameter space. Representative plot for one parameter analyzed on MM34 shown. Four separate chains initialized to different starting parameters are visualized in four colors. The trajectory of the chains are consistent over time. **B)** Gelman plots confirm convergence of chains. Representative plot for one parameter analyzed on MM34 shown. Low chain reduction is stable as seen with the decreased scale reduction factor, suggesting decreasing variance among the four chains with increasing number of iterations as desired. **C)** Distribution of σ^2 (expression variance) is comparable within 100Mb deletion, normal, and amplification regions simulated from known deletion, normal, and amplification regions in MGH31. **D)** Distribution of θ_0 (expected gene expression magnitude shift) for deletions of varying size, simulated by randomly sampled gene sets of varying size within known deletion regions in MM16. Mean of estimate is stable but variable depending on number of genes used for estimation. **E)** Average σ^2 variance around expected gene expression magnitude shift, versus number of genes on a log-log scale. Linear regression fit estimating σ^2 as a function of the number of genes within a putative CNV region is plotted in red.

C. Supplemental References

Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, Ryu HS, Kim S, Lee JE, Park YH, et al. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* **8**: 15081. http://www.nature.com/doifinder/10.1038/ncomms15081 (Accessed May 23, 2017).

Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan J-B, Zhang K, Chun J, et al. 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* **13**: 241–4. http://www.ncbi.nlm.nih.gov/pubmed/26780092 (Accessed August 17, 2016).

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**: 163–6. http://www.ncbi.nlm.nih.gov/pubmed/24363023 (Accessed April 28, 2014).

Kharchenko P V, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11**: 740–2. http://www.ncbi.nlm.nih.gov/pubmed/24836921.

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**: 417–425. http://www.ncbi.nlm.nih.gov/pubmed/26771021 (Accessed August 17, 2016).

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed B V, Curry WT, Martuza RL, et al. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**: 1396–401. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4123637&tool=pmcentrez&rendertype=abstract.